B Some Probability and Information Theory

B.1 Random Variables and Distributions

A **discrete probability space** is given by a countable set Ω and a probability function $P : \Omega \to [0,1]$ with $\sum_{\omega\in\Omega} P(\omega) = 1$. An **event** $\mathcal{A}$ is a subset of Ω, and for any event $\mathcal{A}$ the probability $P[\mathcal{A}]$ of the event is given by $P[\mathcal{A}] := \sum_{\omega\in\mathcal{A}} P(\omega)$. For two events $\mathcal{A}$ and $\mathcal{B}$, the **conditional** probability $P[\mathcal{A}|\mathcal{B}]$ is defined as $P[\mathcal{A}|\mathcal{B}] := P[\mathcal{A}\cap\mathcal{B}]/P[\mathcal{B}]$. A **random variable** is a function $X : \Omega \to \mathcal{X}$. The **distribution** of X is the function $P_X : \mathcal{X} \to [0,1]$ defined as $P_X(x) = P[X{=}x]$, where $X{=}x$ is a shorthand for the event $\{\omega \in \Omega \,|\, X(\omega) = x\}$. Furthermore, we write P_{XY} for the joint distribution of two random variables X and Y, i.e. $P_{XY}(x,y) = P[X = x \wedge Y = y]$, and we write $P_{X|\mathcal{E}}(x) = P[X{=}x|\mathcal{E}]$ and $P_{X|Y}(x|y) = P[X{=}x|Y{=}y]$ for the **conditional distributions** (conditioned on an event $\mathcal{E}$ respectively a random variable Y).

In these lecture notes, we usually leave Ω and $P : \Omega \to [0,1]$ implicit, and understand it as defined by the joint distribution (and probabilities) of all the random variables (and events) involved, where a distribution (with domain $\mathcal{X}$) may be an arbitrary function $Q : \mathcal{X} \to [0,1]$ with $\sum_x Q(x) = 1$. Also, we sometimes abuse notation and write $X \in \mathcal{X}$ to denote that the range of the random variable X is $\mathcal{X}$.

Definition B.1. *The* **statistical distance** *between two distributions P and Q with common domain $\mathcal{X}$ is defined as*

$$\mathrm{SD}(P,Q) := \frac{1}{2}\sum_{x\in\mathcal{X}} \big|P(x) - Q(x)\big|$$

If the distributions describing two experiments have small statistical distance, then this can be interpreted as that the experiments behave in exactly the same way except with small "error" probability. This is formalized by the following lemma.

Lemma B.2. *Let Q and Q' be two probability distributions with common domain $\mathcal{X}$. Then there exists a joint distribution $P_{XX'}$ for random variables X and X' such that $P_X = Q$ and $P_{X'} = Q'$, and such that $P[X \neq X'] = \mathrm{SD}(Q,Q')$.*

B.2 Hoeffding's Inequality

Theorem B.3 (Hoeffding's inequality). *Let $v \in \{0,1\}^n$ be a bit string with relative Hamming weight $\mu = \omega(v) = \sum_i v_i/n$. Let the random variables $X_1, X_2, \ldots, X_k$ be obtained by sampling k random entries from v* with replacement*, i.e., the X_i's are independent and $P_{X_i}(1) = \mu$. Similarly, let $Y_1, Y_2, \ldots, Y_k$ be obtained by sampling k random entries from v* without replacement*. Then, for any $\delta > 0$, the random variables $\bar{X} := \frac{1}{k}\sum_i X_i$ and $\bar{Y} := \frac{1}{k}\sum_i Y_i$ satisfy*

$$\Pr\big[|\bar{Y} - \mu| \geq \delta\big] \leq \Pr\big[|\bar{X} - \mu| \geq \delta\big] \leq 2\exp(-2\delta^2 k)\,.$$

B.3 Jensen's Inequality

Proposition B.4 (Jensen's inequality). *Let $f : I \to \mathbb{R}$ be a convex function on some interval $I \subset \mathbb{R}$. Then, for any $x_1, \ldots, x_n \in I$ and any $0 \leq p_1, \ldots, p_n \in \mathbb{R}$ with $\sum_i p_i = 1$,*

$$\sum_i p_i f(x_i) \geq f\Big(\sum_i p_i x_i\Big)\,.$$

If f is a concave function then the inequality is reversed.

In probability-theoretic terms, Jensen's inequality can be expressed as follows.

Corollary B.5. *Let $f : I \to \mathbb{R}$ be a convex function on some interval $I \subset \mathbb{R}$. Then, for any random variable X over I, it holds that $E[f(X)] \leq f(E[X])$.*

B.4 Shannon Entropy and Mutual Information

Let X and Y be random variables with respective ranges $\mathcal{X}$ and $\mathcal{Y}$. Throughout, "log" denotes the *binary* logarithm.

Definition B.6. *The* **(Shannon) entropy** *of X is defined as*

$$\mathrm{H}(X) := -\sum_x P_X(x) \log P_X(x)\,,$$

where the sum is over all $x \in \mathcal{X}$ with $P_X(x) > 0$.

It is not hard to see that $0 \le \mathrm{H}(X) \le \log|\mathcal{X}|$, with equality on the left if and only if X is constant, i.e. if there is no uncertainty at all in X, and with equality on the right if and only if X is uniform over $\mathcal{X}$, i.e. has maximal uncertainty.

Note that $\mathrm{H}(X)$ is actually a function of the *distribution* P_X of X, and thus we may also write $\mathrm{H}(Q)$ for any distribution Q. Thus, the notion naturally extends to

$$\mathrm{H}(XY) := \mathrm{H}(P_{XY}) = -\sum_{x,y} P_{XY}(x,y) \log P_{XY}(x,y)\,,$$

$$\mathrm{H}(X|Y{=}y) := \mathrm{H}(P_{X|Y=y}) = -\sum_x P_{X|Y}(x|y) \log P_{X|Y}(x|y)\,,$$

etc.

Definition B.7. *The* **conditional (Shannon) entropy** *of X given Y is defined as*

$$\mathrm{H}(X|Y) := \sum_y P_Y(y)\, \mathrm{H}(X|Y=y)\,,$$

where the sum is over all $y \in \mathcal{Y}$ with $P_Y(y) > 0$.

The following rules hold.

Lemma B.8. *For any random variables X, Y and Z:*

1. $\mathrm{H}(XY|Z) \ge \mathrm{H}(X|Z)$ *("more data can only have more uncertainty"),*
2. $\mathrm{H}(X|Z) \ge \mathrm{H}(X|YZ)$ *("side information can only decrease uncertainty"), and*
3. $\mathrm{H}(X|YZ) = \mathrm{H}(XY|Z) - \mathrm{H}(Y|Z)$ *(***chain rule***).*

The "information Y gives on X" is now defined as the loss of entropy in X when Y is given.

Definition B.9. *The* **mutual information** *between X and Y is defined as*

$$I(X;Y) := H(X) - H(X|Y) = H(X) + H(Y) - H(XY) = H(Y) - H(Y|X)\,,$$

and the **conditional mutual information** *between X and Y given Z as*

$$I(X;Y|Z) := H(X|Z) - H(X|YZ)\,.$$

From the above chain rule and the definition of the (conditional) mutual information, we immediately get the chain rule for mutual information:

$$I(X;YZ) = I(X;Z) + I(X;Y|Z)\,.$$

Finally, the mutual information satisfies the following so-called **data-processing** inequality.

Lemma B.10. *If X, Y and Z are random variables, where Z is obtained by processing Y (formally: $X \to Y \to Z$ forms a Markov chain), then $I(X;Z) \leq I(X;Y)$.*

The Shannon entropy and the mutual information have proven to be the right measures for "uncertainty" and "information" in communication theory. For instance, the Shannon entropy captures how far data can be compressed, and the mutual information captures how much information can be reliably communicated over a noisy communication channel.

Definition B.11. *The* **binary entropy function** $\mathrm{h} : [0,1] \to [0,1]$ *is defined as*

$$\mathrm{h}(p) = -\big(p\log(p) + (1-p)\log(1-p)\big)\,,$$

with $\mathrm{h}(0) = 0 = \mathrm{h}(1)$.

Besides expressing the Shannon entropy of a binary random variable, the binary entropy function is also useful for bounding the number of strings with a certain Hamming weight.

Lemma B.12. *The size of the set $B(\alpha n) = \{x \in \{0,1\}^n \,|\, W(x) \leq \alpha n\}$ of n-bit strings with Hamming weight at most αn is upper bounded in size by*

$$|B(\alpha n)| \leq 2^{\mathrm{h}(\alpha)n}\,.$$

B.5 Min- and Collision-Entropy

Here, we define a couple of alternative uncertainty measures. Let X and Y be random variables with respective ranges $\mathcal{X}$ and $\mathcal{Y}$.

Definition B.13. *The* **guessing probability** *and the* **min-entropy** *of X are respectively defined as*

$$\mathrm{Guess}(X) := \max_x P_X(x) \qquad \textit{and} \qquad \mathrm{H}_\infty(X) := -\log\big(\mathrm{Guess}(X)\big) = -\log\big(\max_x P_X(x)\big)\,.$$

Like the Shannon entropy, the min-entropy (and the same will hold for the collision entropy below) is 0 if and only if X is constant, and maximal, i.e., $\log|\mathcal{X}|$ if and only if X is uniform on $\mathcal{X}$, but in-between these two extremes, the min-entropy behaves differently (actually: more conservatively).

Also here, $\mathrm{Guess}(X)$ and $\mathrm{H}_\infty(X)$ are actually functions of the *distribution* P_X of X, and thus we may also write $\mathrm{Guess}(Q)$ and $\mathrm{H}_\infty(Q)$ for any distribution Q, and, as such, $\mathrm{Guess}(XY)$, $\mathrm{H}_\infty(X|Y=y)$, etc. are naturally defined.

Definition B.14. *The* **conditional guessing probability** *and the* **conditional min-entropy** *of X given Y are respectively defined as*

$$\mathrm{Guess}(X|Y) := \sum_y P_Y(y)\mathrm{Guess}(X|Y=y) \qquad \textit{and} \qquad \mathrm{H}_\infty(X|Y) := -\log\big(\mathrm{Guess}(X|Y)\big)\,.$$

Warning: Different notions of *conditional* min-entropy can be found in the literature. The one we are using here is convenient for our purposes.

By replacing the guessing probability by the collision probability, we obtain the notion of (conditional) collision entropy as follows.

Definition B.15. *The* **collision probability** *and the* **collision entropy** *of X are respectively defined as*

$$\mathrm{Col}(X) := \sum_x P_X(x)^2 \qquad \textit{and} \qquad \mathrm{H}_2(X) := -\log\big(\mathrm{Col}(X)\big) = -\log\left(\sum_x P_X(x)^2\right),$$

and the **conditional collision probability** *and* **entropy** *of X given Y are respectively defined as*

$$\mathrm{Col}(X|Y) := \sum_y P_Y(y)\mathrm{Col}(X|Y=y) \qquad \textit{and} \qquad \mathrm{H}_2(X|Y) := -\log\big(\mathrm{Col}(X|Y)\big),$$

where naturally $\mathrm{Col}(X|Y=y) := \sum_x P_{X|Y}(x|y)^2$.

The following rules hold.

Lemma B.16. *For any random variables X, Y and Z:*

1. $\mathrm{H}_\infty(XY|Z) \geq \mathrm{H}_\infty(X|Z)$ *and* $\mathrm{H}_2(XY|Z) \geq \mathrm{H}_2(X|Z)$,
2. $\mathrm{H}_\infty(X|Z) \geq \mathrm{H}_\infty(X|YZ)$ *and* $\mathrm{H}_2(X|Z) \geq \mathrm{H}_2(X|YZ)$, *and*
3. $\mathrm{H}_\infty(X|YZ) \geq \mathrm{H}_\infty(XY|Z) - \log(|\mathcal{Y}|) \geq \mathrm{H}_\infty(X|Z) - \log(|\mathcal{Y}|)$ *(***chain rule** *for min-entropy).*

Finally, the different entropy notions compare to each other as follows.

Lemma B.17. *For any random variables X and Z:* $\mathrm{H}_\infty(X|Z) \leq \mathrm{H}_2(X|Z) \leq \mathrm{H}(X|Z)$.